Configuration and Performance of a Dual Boot (Linux/Tru64) Cluster

Michael Lang, Harvey Wasserman, Darren Kerbyson, Fabrizio Petrini, Adolfy Hoisie, Scott Pakin

Performance and Architecture Team
Modeling, Algorithms and Informatics Group (CCS-3)
Los Alamos National Laboratory

{mlang, hjw, fabrizio, hoisie, djk, pakin}@lanl.gov

LAUR-02-5853

Overview

- Configuration of a dual boot alpha cluster
  - The Hardware
  - Linux Installation
  - Problems
- Reliability
- Performance on LANL codes under both O/S's:
  - Sweep3d
  - SAGE
  - POP

Purpose

To provide a large linux cluster for R & D, while still allowing the team to investigate areas of interest to LANL's Tru64 based Q machine.

To evaluate Linux as a stable environment for LANL Codes.

The Hardware

- Installed in June of 2001 as a Network test-bed for LANL's Q.
- 19 Racks
- 64 Nodes/ 256 Processors
- ES40's, 833MHz processors
- 8GB RAM /node
- Dual Rail Quadrics
- Each node 3-disks (1 for Linux)

The Software

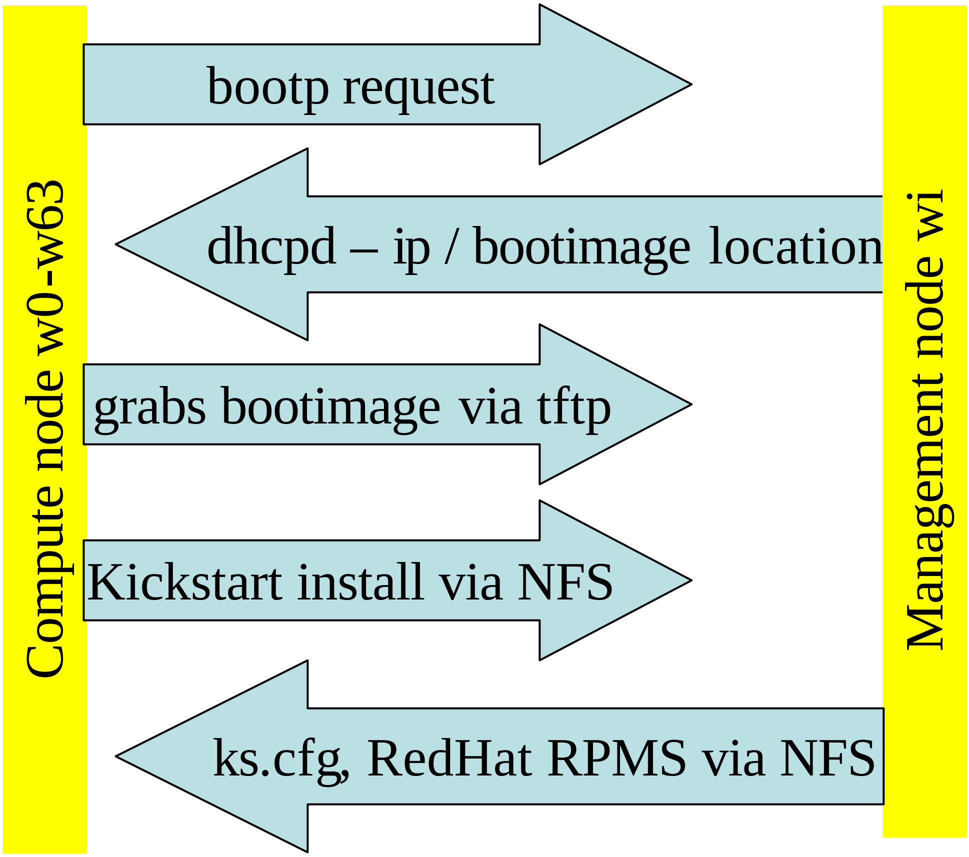
- Sierra Cluster EFT-3
  - Tru64 5.1
- RedHat Linux Version 7.1
  - 2.4.3 Kernel
- Both with Quadrics software.
  - RMS 2.6

Linux Installation

- Nodes installed over the network via bootp & Redhat's Kickstart
- SSI – Single system image(Installed to local disk)

Reinstalled in about 7minutes

All compilers and quadrics configured.



Compute node w0-w63

Management node wi

bootp request

dhcpd – ip / bootimage location

grabs bootimage via tftp

Kickstart install via NFS

ks.cfg, RedHat RPMS via NFS

Linux install cont.

- Use netabootwrap utility to create a kernel/ramdisk for downloadable by bootp.
    - (also has boot options for kickstart installation)
- Uses Redhat's install kernel
- Problems:
    - Had to dd a BSD partition table
    - Utilities are not actively maintained

Console Management

- Perl/Expect scripts were developed to mirror Tru64’s sra commands
  - reboot
  - halt
  - power off/on …

(There is also an opensource project from llnl – conman)

Pros and Cons

- Switching OS’s is fast & easy
  - Reboot management node and start up all the nodes
- Help with trouble shooting hardware software problems

- Two OS’s to update
- No Emulex Linux drivers for our SAN
- Limited software available for alpha-linux
  - KAI C++ compiler; now Intel, now discontinued…

Hardware Reliability

Items Replaced since initial install (~1yr):

~1.6% of components replaced (20/1226)

No such thing as fault tolerant hardware!

Better to have spare nodes automatically configured in as replacements.

This would require fine grained fault tolerance at the OS level.

Component	# Replaced
CPU	3 of 256
Memory Banks	4 of 256
Disks	2 of 192
Fans	2 of 192
PS	2 of 192
PCI Backplane	2 of 64
Motherboard	3 of 64
Quadrics Switchblade	1 of 8
100BT Switch	1 of 2

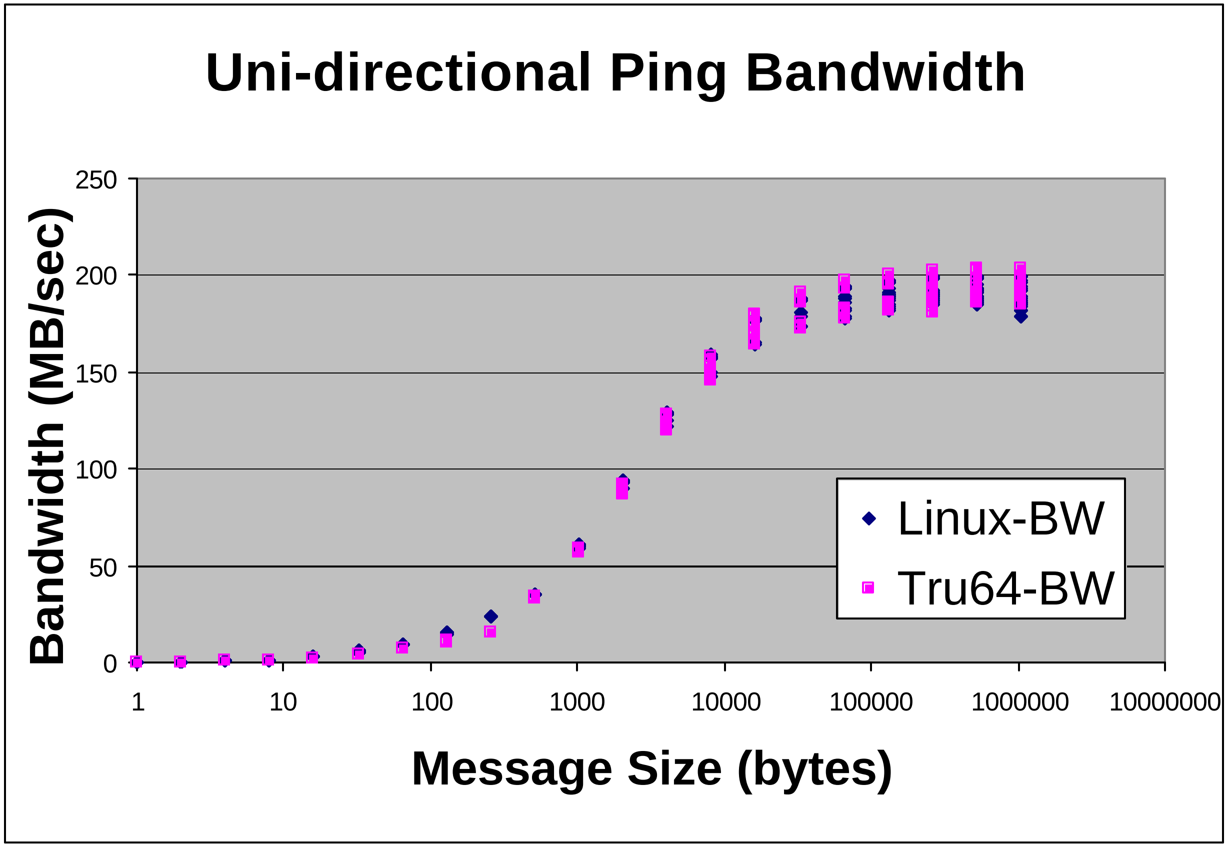
Performance

- Communications
    - Bandwidth and Latency
- Representative workloads of interest to LANL
    - SAGE - Hydrodynamics
    - Sweep3D – Sn Particle using wave propagation
    - POP – Ocean Simulation

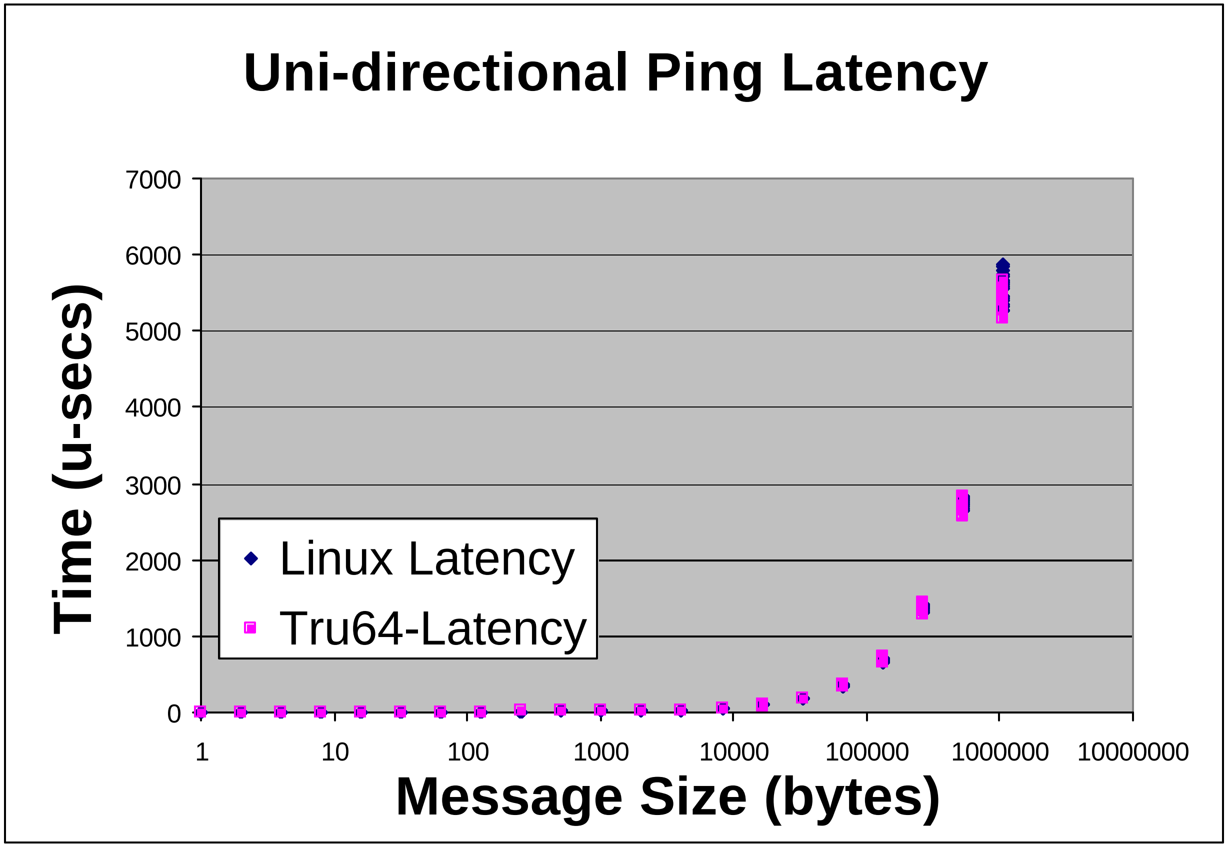
Communication Comparison

- Comms Test Suite:
  - MPI pings
    - Unidirectional
    - Bidirectional
    - Ping all
  - MPI collectives
    - Barrier
    - Broadcast
  - Hotspot

Comms Results



Comms Results cont.



5.5u-sec Latency

Communication Comparison

- Comms Test Suite:
  - MPI pings
    - Unidirectional
    - Bidirectional
    - Ping all
  - MPI collectives
    - Barrier
    - Broadcast
  - Hotspot

No difference Due to OS !!

SAGE

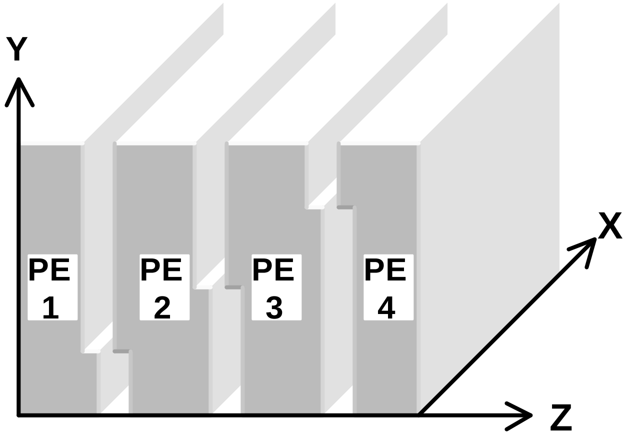
SAGE = SAIC's Adaptive Grid Eulerian, LANL & SAIC.

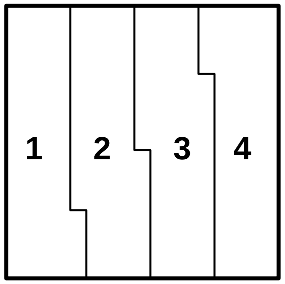
Multidimensional hydrodynamics code with adaptive mesh refinement

- LANL recently modeled a tsunami resulting from an asteroid impact using SAGE. http://www.lanl.gov/orgs/pa/newsbulletin/2002/06/06/text01.shtml

SAGE (cont)

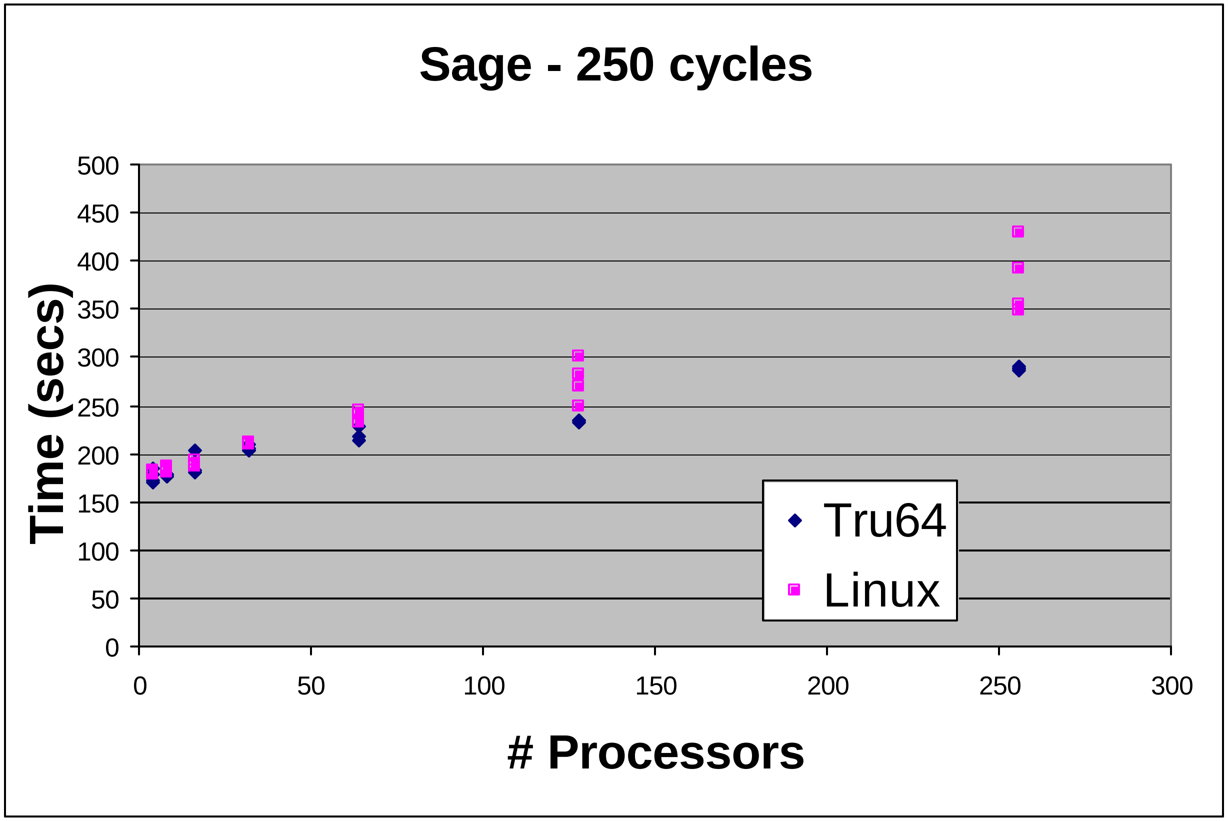
- Parallel Decomposition occurs spatially in sub-grids over the processors. So each PE gets a sub-grid volume to solve. (usually an X,Y slab is assigned to each PE)





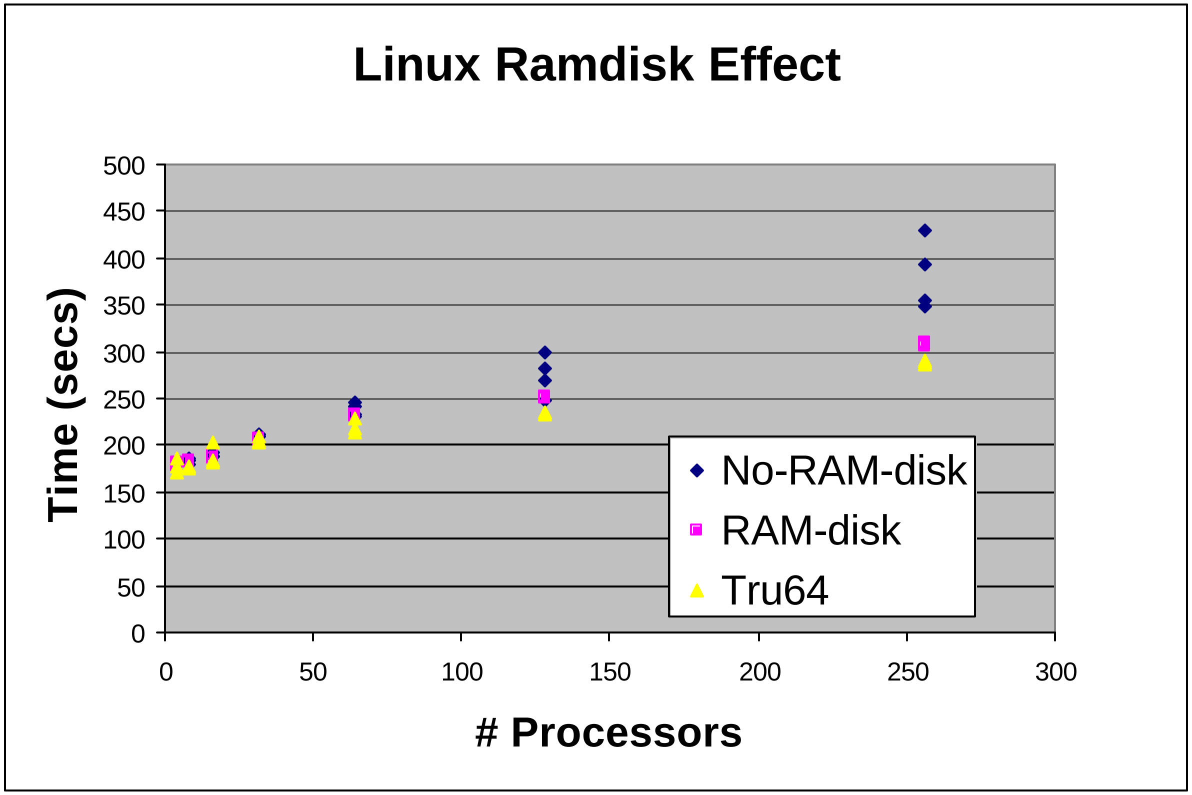
SAGE Performance

- Weak scaling
- Constant amount of work/PE



SAGE Performance cont.

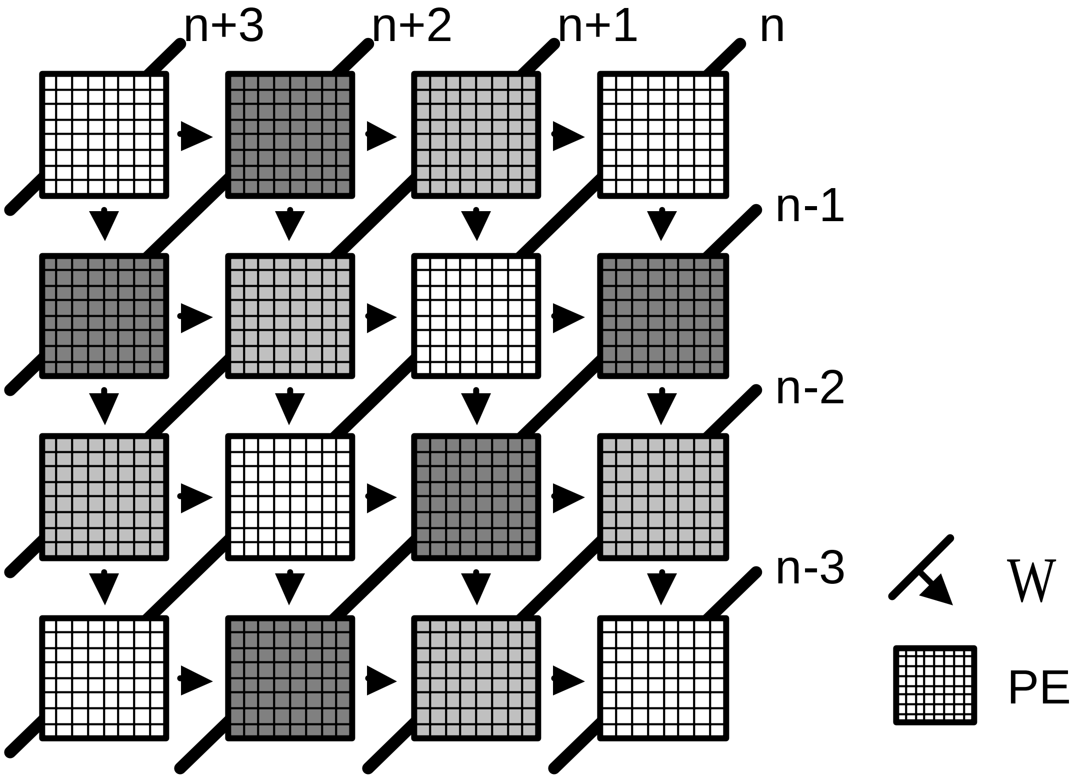
- Difference in high # nodes due to loading the binary from NFS



Sweep

- Sweep3d is an Sn transport kernel.
- 4 dimensions: I, J, K & angle.
- Parallel decomposition occurs in I and J;

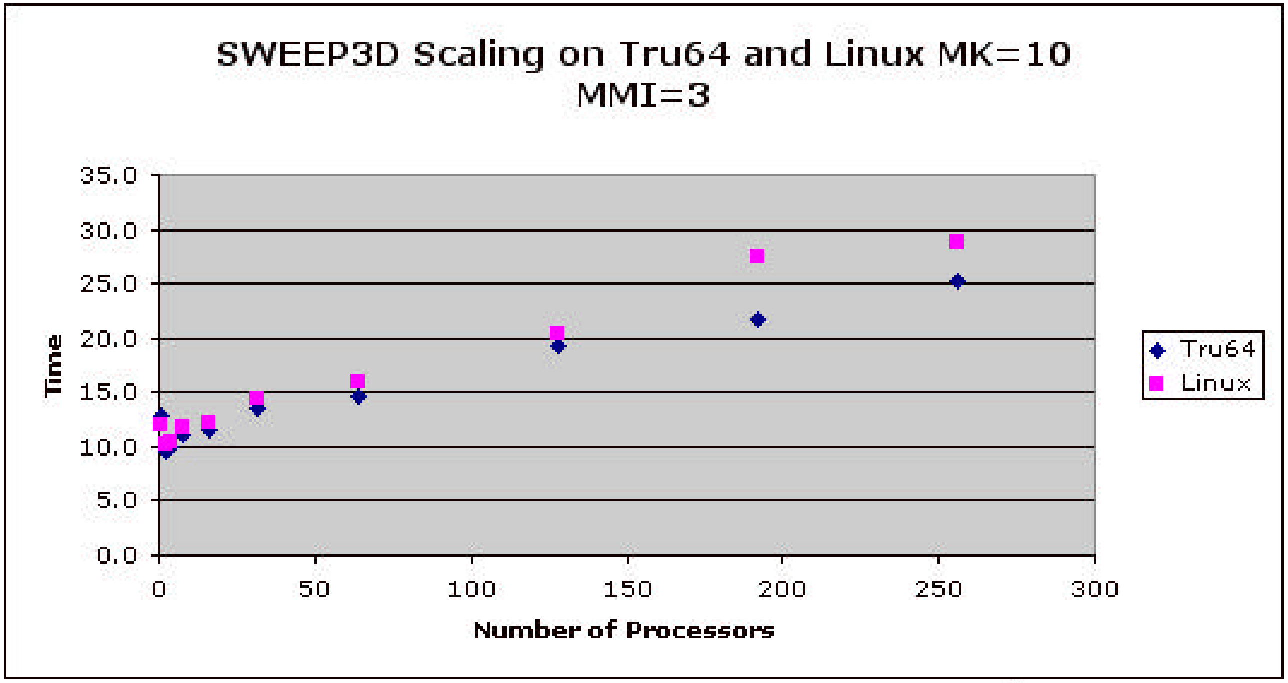
K and angle are solved serially on each node.



Sweep3D Performance

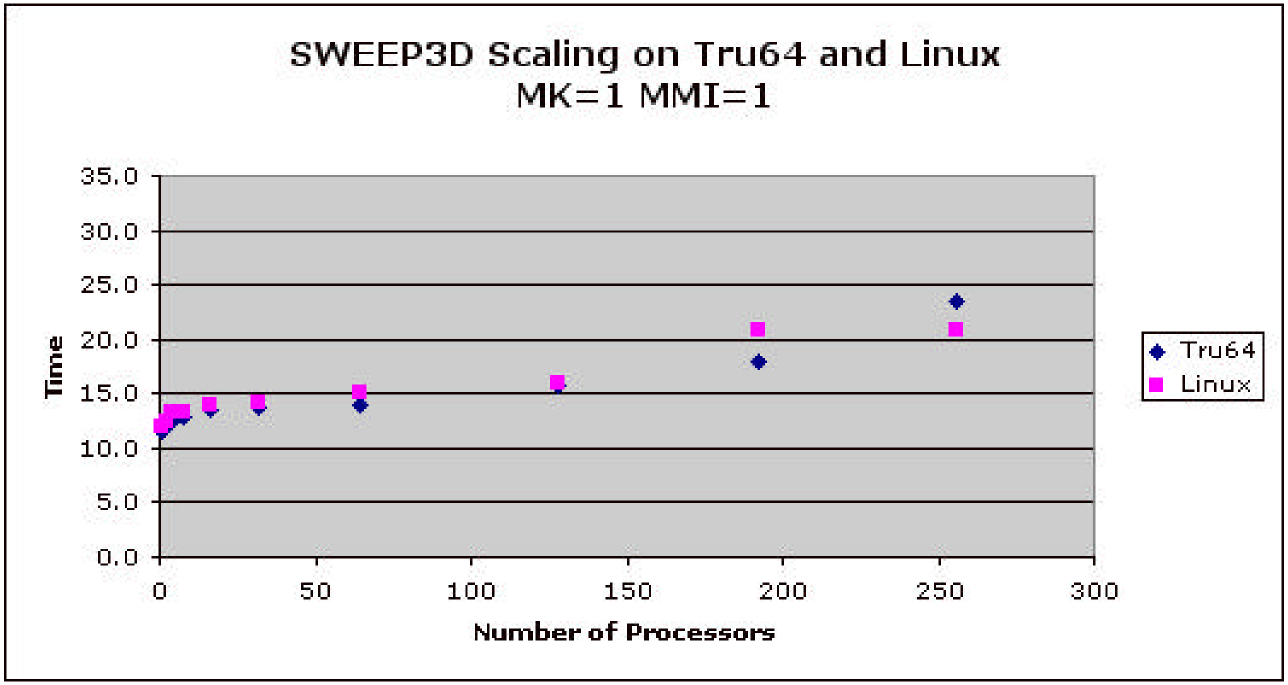
MK is the blocking factor for the K plane
MMI is the blocking factor for the angles

- Weak scaling
- Constant amount of work/PE



Sweep Performance cont.

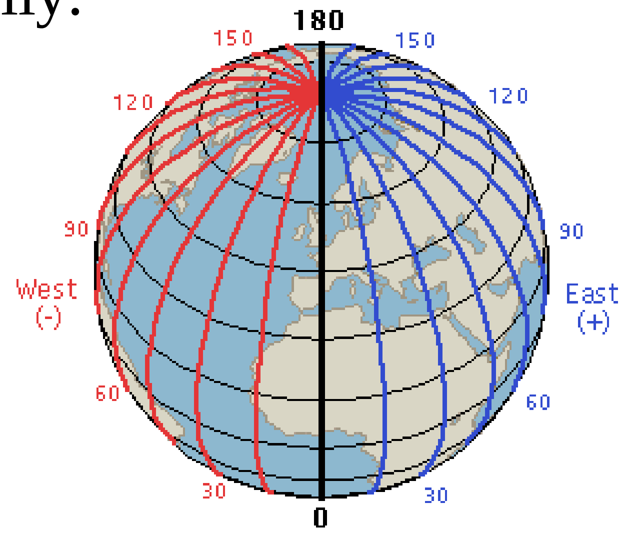
- Results weighted towards communication (smaller pieces)



POP

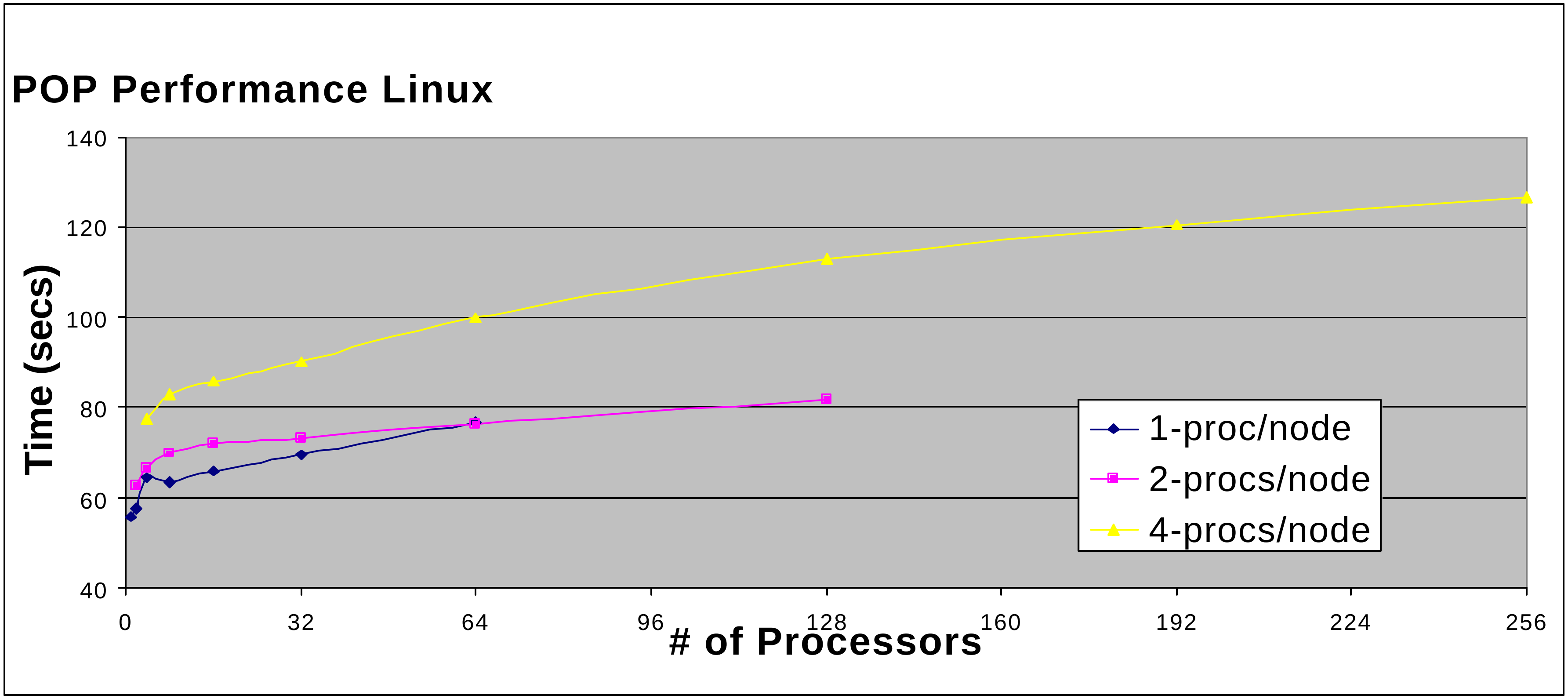
- Ocean Modeling Code
  - www.acl.lanl.gov/climate/models/pop

- Parallel Decomposition using latitude and longitude and with ocean depth solved serially.



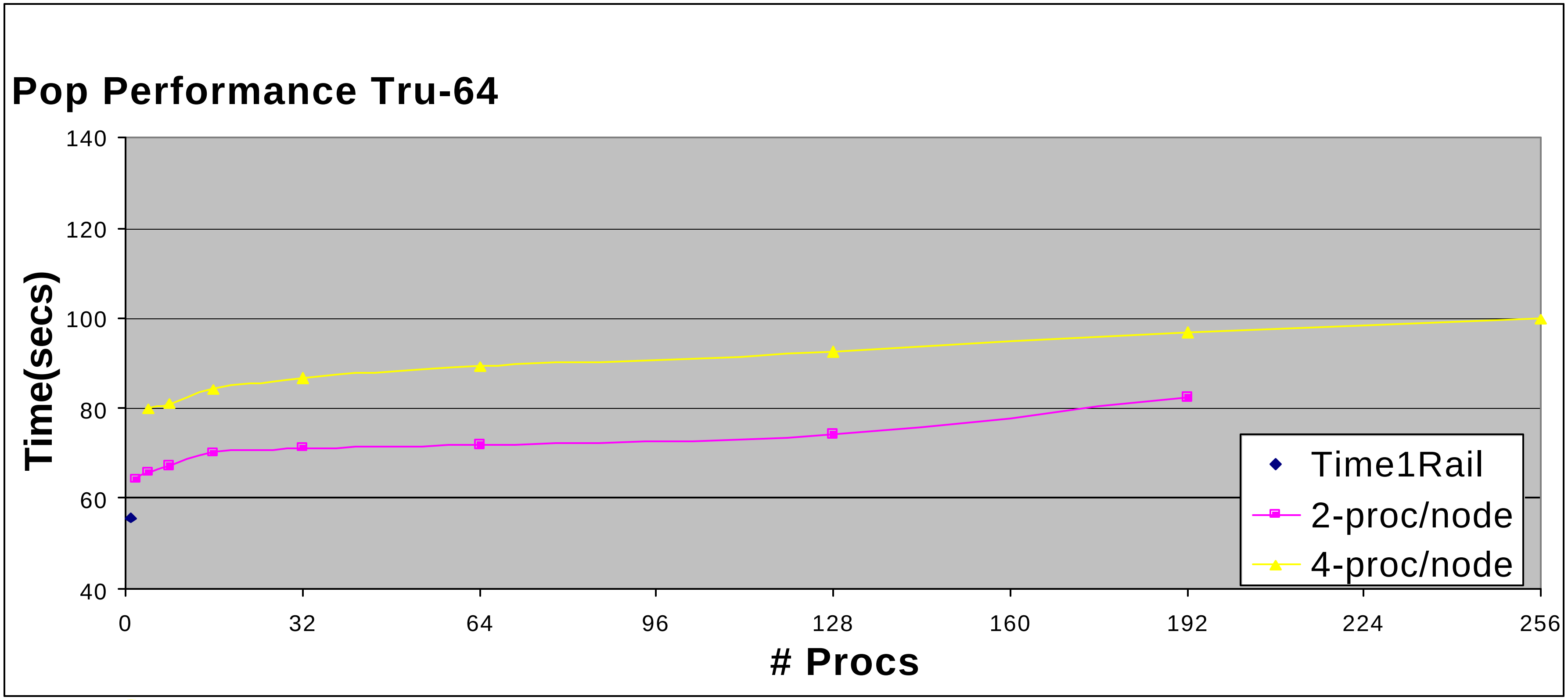
POP Performance

- Weak scaling, Constant amount of work/PE



POP Performance cont.

- Weak scaling, Constant amount of work/PE



Conclusions

- Dual boot cluster is a great resource
  - Switching OS is easy
  - Good for trouble shooting hardware/software
  - Use as production/development environment
- Linux for LANL
  - Which Parallel filesystem? NFS isn't a parallel filesystem!! (Try Netapps, wait for lustre?)
  - Quality Linux compilers (Intel??)
- Revisit with new version of SC and Linux

and Quadrics software.

Thanks!!

- ASCI
- CCS-3 Performance and Architecture Team
- HP (hardware support team @ LANL)
- Quadrics

Links

These slides	http://www.c3.lanl.gov/~mlang/CAST.html
Storm – resource management	http://www.c3.lanl.gov/~fabrizio/papers/sc02.pdf
Modeling/Scaling of SAGE	http://www.sc2001.org/papers/pap.pap255.pdf
Sweep3d	http://public.lanl.gov/hjw/CODES/SWEEP3D/sweep3d.html
POP	http://www.acl.lanl.gov/climate/models/pop
Conman	http://www.lnll.gov/linux/conman.html